本节可配合第十四讲观看

因子分析模型

因子分析由斯皮尔曼在1904年首次提出，其在某种程度上可以被看成是主成分分析的推广和扩展。

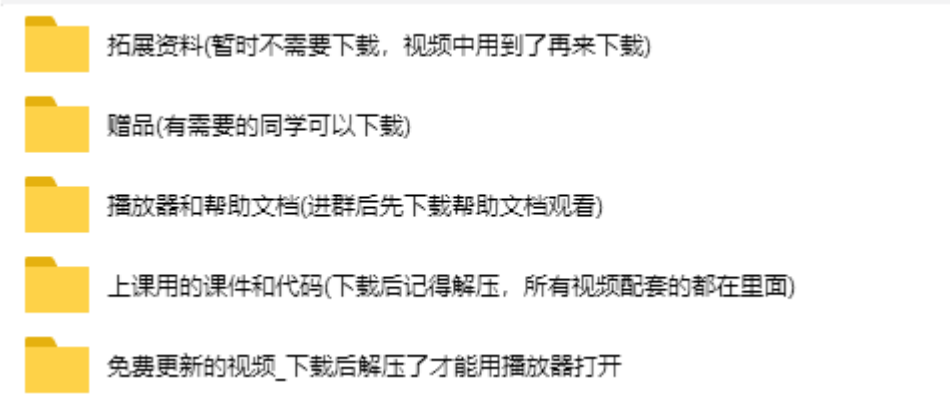
因子分析法通过研究变量间的相关系数矩阵，把这些变量间错综复杂的关系归结成少数几个综合因子，由于归结出的因子个数少于原始变量的个数，但是它们又包含原始变量的信息，所以，这一分析过程也称为降维。由于因子往往比主成分更易得到解释，故因子分析比主成分分析更容易成功，从而有更广泛的应用。

本讲的前面部分将简要介绍因子分析模型的数学原理，在最后的应用部分，我们将举一个实例帮助大家理解，大家可以把重点放在最后的应用上。

温馨提示

（1）视频中提到的附件可在**售后群的群文件**中下载。

包括**讲义、代码、我视频中推荐的资料**等。



（2）关注我的**微信公众号《数学建模学习交流》**，后台发送**“软件”**两个字，可获得常见的建模软件下载方法；发送**“数据”**两个字，可获得建模数据的获取方法；发送**“画图”**两个字，可获得数学建模中常见的画图方法。另外，也可以看看公众号的历史文章，里面发布的都是对大家有帮助的技巧。

（3）**购买更多优质精选的数学建模资料**，可关注我的微信公众号《数学建模学习交流》，在后台发送**“买”**这个字即可进入店铺进行购买。

（4）视频价格不贵，但价值很高。单人购买观看只需要**58元**，和另外两名队友一起购买人均仅需**46元**，视频本身也是下载到本地观看的，所以请大家**不要侵犯知识产权**，对视频或者资料进行二次销售。

数学建模学习交流

因子分析和主成分分析的对比

假设有n个样本，p个指标，则可构成大小为$n\times p$的样本矩阵 $x=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}=(x_1,x_2,\cdots,x_p)$

主成分分析：$x_1,x_2,\cdots,x_p \Rightarrow z_1,z_2,\cdots,z_m(m\le p)$，且它们满足：$\begin{cases} z_1=l_{11}x_1+l_{12}x_2+\cdots+l_{1p}x_p \\ z_2=l_{21}x_1+l_{22}x_2+\cdots+l_{2p}x_p \\ \quad\vdots \\ z_m=l_{m1}x_1+l_{m2}x_2+\cdots+l_{mp}x_p \end{cases}$

$z_1,z_2,\cdots,z_m$是m个主成分，可以看出，主成分实际上就是各指标的线性组合。

因子分析：$x_1,x_2,\cdots,x_p \Rightarrow f_1,f_2,\cdots,f_m(m\le p)$，且它们满足：$\begin{cases} x_1=u_1+a_{11}f_1+a_{12}f_2+\cdots+a_{1m}f_m+\varepsilon_1 \\ x_2=u_2+a_{21}f_1+a_{22}f_2+\cdots+a_{2m}f_m+\varepsilon_2 \\ \quad\vdots \\ x_p=u_p+a_{p1}f_1+a_{p2}f_2+\cdots+a_{pm}f_m+\varepsilon_p \end{cases}$

$f_1,f_2,\cdots,f_m$被称为公共因子，ε_i为特殊因子，各因子的线性组合构成了原始的指标。

（有点像回归，回归中自变量是已知的，因子分析是只知道因变量，要我们来找自变量）

其他主要区别:
1.主成分分析只是简单的数值计算，不需要构造一个模型，几乎没什么假定；而因子分析需要构造一个因子模型，并伴随几个关键性的假定。
2.主成分的解是唯一的，而因子可有许多解。
因子解释成功的可能性要远大于主成分解释成功的可能性。

因子分析的实例

来自：MOOC上王学民老师的多元统计分析

【例1】林登(Linden)根据他收集的来自139名运动员的比赛数据，对第二次世界大战以来奥林匹克十项全能比赛的得分作了因子分析研究。这十个全能项目为：100米跑(x_1)，跳远(x_2)，铅球(x_3)，跳高(x_4)，400米跑(x_5)，11米跨栏(x_6)，铁饼(x_7)，撑杆跳高(x_8)，标枪(x_9)，1500米跑(x_{10})。经标准化后所作的因子分析表明，十项得分基本上可归结于他们的短跑速度、爆发性臂力、爆发性腿力和耐力这四个方面，每一方面都称为一个因子。十项得分与这四个因子之间的关系可以描述为如下的因子模型：

$$x_i=\mu_i+a_{i1}f_1+a_{i2}f_2+a_{i3}f_3+a_{i4}f_4+\varepsilon_i,\ \ i=1,2,\cdots,10$$

其中f_1,f_2,f_3,f_4表示四个因子，称为公共因子(common factor)，a_{ij}称为x_i在因子f_j上的载荷(loading)，μ_i是x_i的均值，ε_i是x_i不能被四个公共因子解释的部分，称之为特殊因子(specific factor)。

因子分析的实例

来自：MOOC上王学民老师的多元统计分析

【例2】公司老板对48名应聘者进行面试，并给出他们在15个方面所得的分数，这15个方面是：

x_1：申请书的形式
x_2：外貌
x_3：专业能力
x_4：讨人喜欢
x_5：自信心
x_6：精明
x_7：诚实
x_8：推销能力
x_9：经验
x_{10}：积极性
x_{11}：抱负
x_{12}：理解能力
x_{13}：潜力
x_{14}：交际能力
x_{15}：适应性

通过因子分析，这15个方面可以归结为应聘者的社交能力、经验、讨人喜欢的程度、专业能力和外貌这五个因子。

因子分析的原理

假设p维随机向量 $x=(x_1,x_2,\cdots,x_p)'$的均值$u=(u_1,u_2,\cdots u_p)'$,协方差矩阵$\Sigma_{p\times p}=(\sigma_{ij})$

因子分析的一般模型为：
$$\begin{cases} x_1=u_1+a_{11}f_1+a_{12}f_2+\cdots+a_{1m}f_m+\varepsilon_1 \\ x_2=u_2+a_{21}f_1+a_{22}f_2+\cdots+a_{2m}f_m+\varepsilon_2 \\ \quad\vdots \\ x_p=u_p+a_{p1}f_1+a_{p2}f_2+\cdots+a_{pm}f_m+\varepsilon_p \end{cases}$$

其中$f_1,f_2,\cdots,f_m$被称为公共因子，$\varepsilon_i(i=1,2,\cdots,p)$为特殊因子，它们都是无法观测的随机变量。公因子$f_1,f_2,\cdots,f_m$出现在每一个原始变量$x_i(i=1,2,\cdots,p)$的表达式中，可以理解为原始变量共同拥有的某些特征（具有共同的影响因素）；每个特殊因子$\varepsilon_i(i=1,2,\cdots,p)$仅仅出现在与之相应的第$i$个原始变量$x_i$的表达式中，它只对这个原始变量起作用。

上面这个式子我们用矩阵形式可记为：$x=u+Af+\varepsilon$

其中$f=(f_1,f_2,\cdots,f_m)'(m\le p)$为公因子向量，$\varepsilon=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_p)'$为特殊因子向量，$A_{p\times m}=(a_{ij})$称为因子载荷矩阵，并假设$A$的秩为$m$.

要进行因子分析，必须要解出A这个矩阵，因此下面我们要给的一些假设用来计算A矩阵。

因子分析模型的假设

$$\begin{cases} x_1 = u_1 + a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m + \varepsilon_1 \\ x_2 = u_2 + a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + \varepsilon_2 \\ \quad \vdots \\ x_p = u_p + a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m + \varepsilon_p \end{cases}$$

$$x = u + Af + \varepsilon，\text{假设} \begin{cases} E(f) = 0 \\ E(\varepsilon) = 0 \\ Var(f) = I \\ Var(\boldsymbol{\varepsilon}) = D = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2) \\ cov(f, \boldsymbol{\varepsilon}) = E(f\boldsymbol{\varepsilon}') = 0 \end{cases}$$

其中$f = (f_1, f_2, \cdots, f_m)'(m \le p)$为公因子向量

$\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p)'$为特殊因子向量

$A_{p\times m} = (a_{ij})$称为因子载荷矩阵，并假设A的秩为m.

公因子彼此不相关，且具有单位方差；特殊因子彼此不相关且与公因子也不相关。

因子模型的性质

（1）x的协方差矩阵Σ的分解

$$x=u+Af+\varepsilon\ ,\ \text{假设}\begin{cases}E(f)=0\\E(\varepsilon)=0\\Var(f)=I\\Var(\boldsymbol{\varepsilon})=D=diag(\sigma_1^2,\sigma_2^2,\cdots,\sigma_p^2)\\cov(f,\boldsymbol{\varepsilon})=E(f\boldsymbol{\varepsilon}')=0\end{cases}$$

$$\begin{aligned}Var(x)\ &=\ E[(x-u)(x-u)']\ =\ E[(Af+\varepsilon)(Af+\varepsilon)']\\&=\ AE(ff')A'\ +\ AE(f\varepsilon')\ +\ E(\varepsilon f')A'\ +\ E(\varepsilon\varepsilon')\\&=\ AVar(f)A'+\ Var(\boldsymbol{\varepsilon})\\&=\ AA'\ +\ D\end{aligned}$$

正是因为因子载荷矩阵A不是唯一的，在实际的应用中我们常常利用这一点，通过因子的变换，使得新的因子具有更容易解释的实际意义。

这就是因子分析往往比主成分分析的结果更容易解释的原因。

（2）因子载荷不唯一

令T为任意一个$m\times m$的正交矩阵，令$A^*=AT,\ f^*\ =\ T'f$，则模型可表示为：

$$x=u+A^*f^*+\varepsilon,\ \text{因为假设仍然成立}\begin{cases}E(f^*)=T'E(f)=0\\E(\varepsilon)=0\\Var(f^*)=T'Var(f)T\ =\ T'IT=I\\Var(\boldsymbol{\varepsilon})=D=diag(\sigma_1^2,\sigma_2^2,\cdots,\sigma_p^2)\\cov(f^*,\boldsymbol{\varepsilon})=E(f^*\boldsymbol{\varepsilon}')=T'E(f\boldsymbol{\varepsilon}')\ =\ 0\end{cases}$$

因子载荷矩阵的统计意义

$$\begin{cases} x_1 = u_1 + a_{11} f_1 + a_{12} f_2 + \cdots + a_{1m} f_m + \varepsilon_1 \\ x_2 = u_2 + a_{21} f_1 + a_{22} f_2 + \cdots + a_{2m} f_m + \varepsilon_2 \\ \quad \vdots \\ x_p = u_p + a_{p1} f_1 + a_{p2} f_2 + \cdots + a_{pm} f_m + \varepsilon_p \end{cases}$$

$x = u + Af + \varepsilon$，其中$f = (f_1, f_2, \cdots, f_m)'(m \le p)$为公因子向量，$\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p)'$为特殊因子向量

$A_{p\times m} = (a_{ij})$称为因子载荷矩阵，并假设A的秩为m.

（1）A的元素a_{ij}：原始变量x_i与公因子f_j之间的协方差：$a_{ij} = cov(x_i, f_j)$

如果x经过了标准化，则$a_{ij} = \rho(x_i, f_j)$ (x_i和f_j的相关系数)

（2）A的行元素平方和$h_i^2 = \sum_{j=1}^{m} a_{ij}^2$：原始变量$x_i$对公因子依赖的程度

可以证明：$Var(x_i) = h_i^2 + \sigma_i^2 \ (i = 1, 2, \cdots, p)$

h_i^2反应了公因子对于x_i的影响，可以看成是公因子对于x_i的方差贡献，称为共性方差；而σ_i^2是特殊因子ε_i对x_i的方差贡献，称为个性方差。如果x经过了标准化，则$h_i^2 + \sigma_i^2 = 1$.

因子载荷矩阵的统计意义

$$\begin{cases} x_1 = u_1 + a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m + \varepsilon_1 \\ x_2 = u_2 + a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + \varepsilon_2 \\ \quad \vdots \\ x_p = u_p + a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m + \varepsilon_p \end{cases}$$

$x = u + Af + \varepsilon$，其中$f = (f_1, f_2, \cdots, f_m)'(m \le p)$为公因子向量，$\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p)'$为特殊因子向量

$A_{p\times m} = (a_{ij})$称为因子载荷矩阵，并假设A的秩为m.

（3）A的列元素平方和$g_j^2 = \sum_{i=1}^{p} a_{ij}^2$：公因子$f_j$对$x$的贡献

可以证明：$\sum_{i=1}^{p} Var(x_i) = \sum_{i=1}^{p} a_{i1}^2 Var(f_1) + \sum_{i=1}^{p} a_{i2}^2 Var(f_2) + \cdots + \sum_{i=1}^{p} a_{im}^2 Var(f_m) + \sum_{i=1}^{p} Var(\varepsilon_i)$

$$= g_1^2 Var(f_1) + g_2^2 Var(f_2) + \cdots + g_m^2 Var(f_m) + \sum_{i=1}^{p} \sigma_i^2$$

$$= g_1^2 + g_2^2 + \cdots + g_m^2 + \sum_{i=1}^{p} \sigma_i^2$$

从上述的推导中可以看出，A的第j列元素的平方和g_j^2是$Var(f_j)$的系数，g_j^2的值越大，反映了f_j对x的影响越大，g_j^2是衡量公因子f_j重要性的一个尺度，可视为公因子f_j对x的贡献。

参数估计

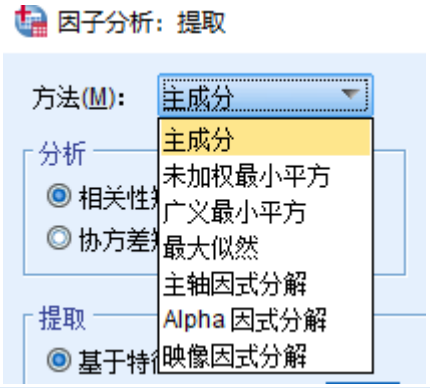
设$x_1, x_2, \cdots, x_n$是一组p维样本，则u和Σ可分别估计为：$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ 和 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$

为了建立因子模型，我们需要估计出因子载荷矩阵$A_{p\times m} = (a_{ij})$，以及个性方差矩阵$D = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)$.

SPSS中提供的方法有主成分法、未加权的最小平方法、综合最小平方法、最大似然法、主轴因子法、Alpha因式分解法和映像因子法。

- **主成分法**，假设变量是因子的线性组合，第一主成分有最大的方差，后续主成分所解释的方差逐渐减小，各主成分之间互不相关，主成分法通常用来计算初始公因子，它也适用于相关矩阵为奇异时的情况。
- **未加权最小平方法**，使得观测的相关矩阵和再生的相关矩阵之差的平方和最小，忽略对角元素。
- **综合最小平方法**，使得观测的相关矩阵和再生的相关矩阵之差的平方和最小，并以变量单值的倒数对相关系数加权。
- **最大似然法**，假设样本来自多元正态分布，使用极大使然估计。
- **主轴因子法**，从初始相关矩阵提取公共因子，并把多元相关系数的平方置于对角线上，再用初始因子载荷估计新的变量共同度，如此重复直至变量共同度在两次相邻迭代中的变化达到临界条件。
- **Alpha 因子法**，把当前分析变量看作是所有潜在变量的一个样本，最大化因子的Alpha可靠性。
- **映像因子法**，把每个变量的主要部分定义为其他各变量的线性回归，而不是潜在因子的函数。

各方法的介绍来自：Spss官方文档



具体选择哪种没有严格的规定，在实际操作中大家可以选择方便最后解释的。论文中最常用主成分法、最大似然法和主轴因子法。

因子旋转的方法

得到因子模型后，其中的公共因子不一定能反映问题的实质特征，为了能更好地解释每一个公共因子的实际意义，且减少解释的主观性，可以通过因子旋转达到目的。

因子旋转分为正交旋转与斜交旋转，经过正交旋转而得到的新的公共因子仍然保持彼此独立的性质，而斜交旋转得到的公共因子是相关的(违背了最初的假定，因此可以看作传统因子分析的拓展)，其实际意义更容易解释。但**不论是正交旋转还是斜交旋转，都应当使新公共因子的载荷系数的绝对值尽可能接近0或1（这里默认了我们从相关系数矩阵进行计算）。**

$$\begin{cases} x_1 = u_1 + a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m + \varepsilon_1 \\ x_2 = u_2 + a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + \varepsilon_2 \\ \quad \vdots \\ x_p = u_p + a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m + \varepsilon_p \end{cases}$$

$x = u + Af + \varepsilon$，其中$f = (f_1, f_2, \cdots, f_m)'(m \le p)$为公因子向量，$\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p)'$为特殊因子向量

$A_{p\times m} = (a_{ij})$称为因子载荷矩阵，并假设A的秩为m.

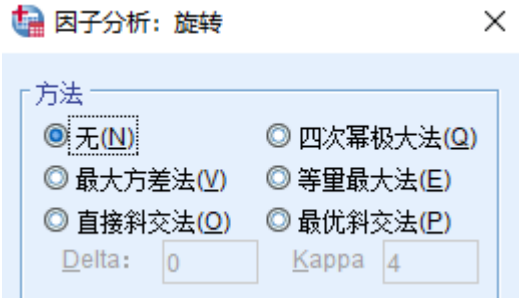
（1）A的元素a_{ij}：原始变量x_i与公因子f_j之间的协方差：$a_{ij} = cov(x_i, f_j)$

如果x经过了标准化，则$a_{ij} = \rho(x_i, f_j)$ (x_i和f_j的相关系数)

SPSS中因子旋转的方法

方法。使您可以选择因子旋转的方法。可用的方法有最大方差、直接 Oblimin、最大四次方值、最大平衡值或最优斜交。

- *最大方差法 (Varimax Method)*. 一种正交旋转方法，它使得对每个因子有高负载的变量的数目达到最小。该方法简化了因子的解释。
- *直接 Oblimin 方法*。一种斜交（非正交）旋转方法。当 delta 等于 0（缺省值）时，解是最斜交的。delta 负得越厉害，因子的斜交度越低。要覆盖缺省的 delta 值 0，请输入小于等于 0.8 的数。
- *最大四次方值法 (Quartimax Method)*. 一种旋转方法，它可使得解释每个变量所需的因子最少。该方法简化了观察到的变量的解释。
- *最大平衡值法 (Equamax Method)*. 一种旋转方法，它是简化因子的最大方差法与简化变量的最大四次方值法的组合。它可以使得高度依赖因子的变量的个数以及解释变量所需的因子的个数最少。
- *最优斜交旋转 (Promax Rotation)*. 斜交旋转，可使因子相关联。该旋转可比直接最小斜交旋转更快地计算出来，因此适用于大型数据集。



具体选择哪种没有严格的规定，在实际操作中大家可以选择方便最后解释的。
在论文中，使用最多的就是最大方差法。

因子得分

因子分析是将变量表示为公共因子和特殊因子的线性组合；此外，我们可以反过来将公共因子表示为原变量的线性组合，即可得到因子得分。

$$\begin{cases} x_1 = u_1 + a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m + \varepsilon_1 \\ x_2 = u_2 + a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + \varepsilon_2 \\ \quad \vdots \\ x_p = u_p + a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m + \varepsilon_p \end{cases} \Rightarrow \begin{cases} f_1 = b_{11}x_1 + b_{12}x_2 + \cdots + b_{1p}x_p \\ f_2 = b_{21}x_1 + b_{22}x_2 + \cdots + b_{2p}x_p \\ \quad \vdots \\ f_m = b_{m1}x_1 + b_{m2}x_2 + \cdots + b_{mp}x_p \end{cases}$$

第i个因子的得分可写成$f_i = b_{i1}x_1 + b_{i2}x_2 + \cdots + b_{ip}x_p \quad (i = 1, 2, \cdots, m)$

b_{ij}就是第i个因子的得分对应于第j个变量x_j的系数

注：我们计算出因子得分函数的系数后，就能够求出所有的因子得分。

方法。计算因子得分的可选方法有回归、Bartlett 和 Anderson-Rubin。

- *回归法 (Regression Method)*. 一种估计因子得分系数的方法。生成的分数的平均值为 0，方差等于估计的因子分数和真正的因子值之间的平方多相关性。即使因子是正交的，分数也可能相关。
- *Bartlett 得分*。一种估计因子得分系数的方法。所产生分数的平均值为 0。使整个变量范围中所有唯一因子的平方和达到最小。 巴特莱特因子得分
- *Anderson-Rubin 方法 (Anderson-Rubin Method)*. 一种估计因子得分系数的方法；它对 Bartlett 方法做了修正，从而确保被估计的因子的正交性。生成的分数平均值为 0，标准差为 1，且不相关。 安德森—鲁宾因子得分法

注：论文中最常用第三种方法。

因子分析的实例

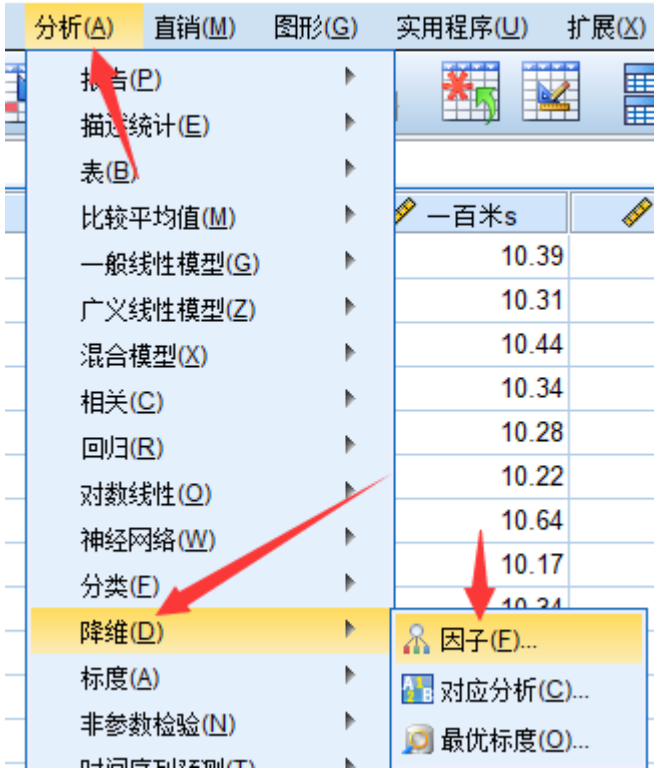
例1：在1984年洛杉矶奥运会田径统计手册中，有55个国家和地区的如下八项男子径赛运动记录：

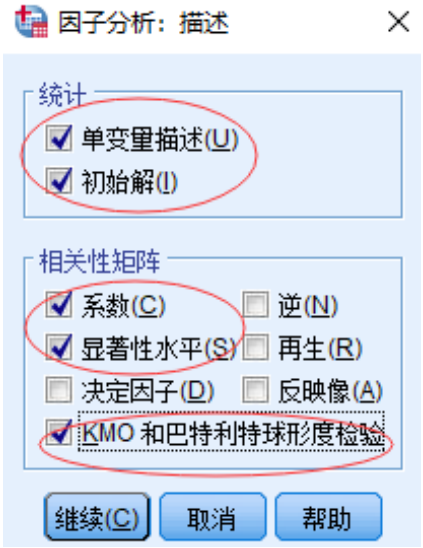
X1: 100米（单位：秒）　　x5: 1500米（单位：分）
x2: 200米（单位：秒）　　x6: 5000米（单位：分）
x3: 400米（单位：秒）　　x7: 10000米（单位：分）
x4: 800米（单位：秒）　　x8: 马拉松（单位：分）

请对该数据进行因子分析。

序号	国家	一百米s	两百米s	四百米s	八百米min	一千五百米min	五千米min	一万米min	马拉松min
1	阿根廷	10.39	20.81	46.84	1.81	3.70	14.04	29.36	137.72
2	澳大利亚	10.31	20.06	44.84	1.74	3.57	13.28	27.66	128.30
3	奥地利	10.44	20.81	46.82	1.79	3.60	13.26	27.72	135.90
4	比利时	10.34	20.68	45.04	1.73	3.60	13.22	27.45	129.95
5	百慕大	10.28	20.58	45.91	1.80	3.75	14.68	30.55	146.62
6	巴西	10.22	20.43	45.21	1.73	3.66	13.62	28.62	133.13
7	缅甸	10.64	21.52	48.30	1.80	3.85	14.45	30.28	139.95
8	加拿大	10.17	20.22	45.68	1.76	3.63	13.55	28.09	130.15
9	智利	10.34	20.80	46.20	1.79	3.71	13.61	29.30	134.03
10	中国	10.51	21.04	47.30	1.81	3.73	13.90	29.13	133.53
11	哥伦比亚	10.43	21.05	46.10	1.82	3.74	13.49	27.88	131.35
12	库克群岛	12.18	23.20	52.94	2.02	4.24	16.70	35.38	164.70
13	哥斯达黎加	10.94	21.90	48.66	1.87	3.84	14.03	28.81	136.58
14	捷克斯洛伐克	10.35	20.65	45.64	1.76	3.58	13.42	28.19	134.32
15	丹麦	10.56	20.52	45.89	1.78	3.61	13.50	28.11	130.78
16	多米尼加共和国	10.14	20.65	46.80	1.82	3.82	14.91	31.45	154.12
17	芬兰	10.43	20.69	45.49	1.74	3.61	13.27	27.52	130.87
18	法国	10.11	20.38	45.28	1.73	3.57	13.34	27.97	132.30
19	德意志民主共和国	10.12	20.33	44.87	1.73	3.56	13.17	27.42	129.92
20	德意志联邦共和国	10.16	20.37	44.50	1.73	3.53	13.21	27.61	132.23
21	大不列颠及北爱尔兰	10.11	20.21	44.93	1.70	3.51	13.01	27.51	129.13
22	希腊	10.22	20.71	46.56	1.78	3.64	14.59	28.45	134.60
23	危地马拉	10.98	21.82	48.40	1.89	3.80	14.16	30.11	139.33

操作步骤





注：我使用的版本为Spss24，没安装的同学可以关注我的微信公众号“数学建模学习交流”，在后台发送“软件”即可。

因子分析：统计

统计

- **单变量描述：**输出参与分析的每个原始变量的均值、标准差和有效取值个数。
- **初始解：**输出未经过旋转直接计算得到的初始公因子、初始特征值和初始方差贡献率等信息。

相关性矩阵

- **系数：**输出初始分析变量间的相关系数矩阵。
- **显著性水平：**输出每个相关系数对于单侧假设检验的显著性水平。
- **决定因子：**输出相关系数矩阵的行列式。
- **逆：**输出相关系数的逆矩阵。
- **再生：**输出因子分析后的相关矩阵，还给出原始相关与再生相关之间的差值，即残差。
- **反映像：**输出反映像相关矩阵，包括偏相关系数的负数。
- **KMO检验和巴特利特球形检验：**进行因子分析前要对数据进行KMO检验和巴特利特球形检验。

KMO检验和巴特利特球形检验

KMO检验

KMO检验是 Kaiser, Meyer和 Olkin提出的，该检验是对原始变量之间的简单相关系数和偏相关系数的相对大小进行检验，主要应用于多元统计的因子分析。

KMO统计量是取值在0和1之间，当所有变量间的简单相关系数平方和远远大于偏相关系数平方和时，KMO值越接近于1，意味着变量间的相关性越强，原有变量越适合作因子分析；当所有变量间的简单相关系数平方和接近0时，KMO值越接近于0,意味着变量间的相关性越弱，原有变量越不适合作因子分析。

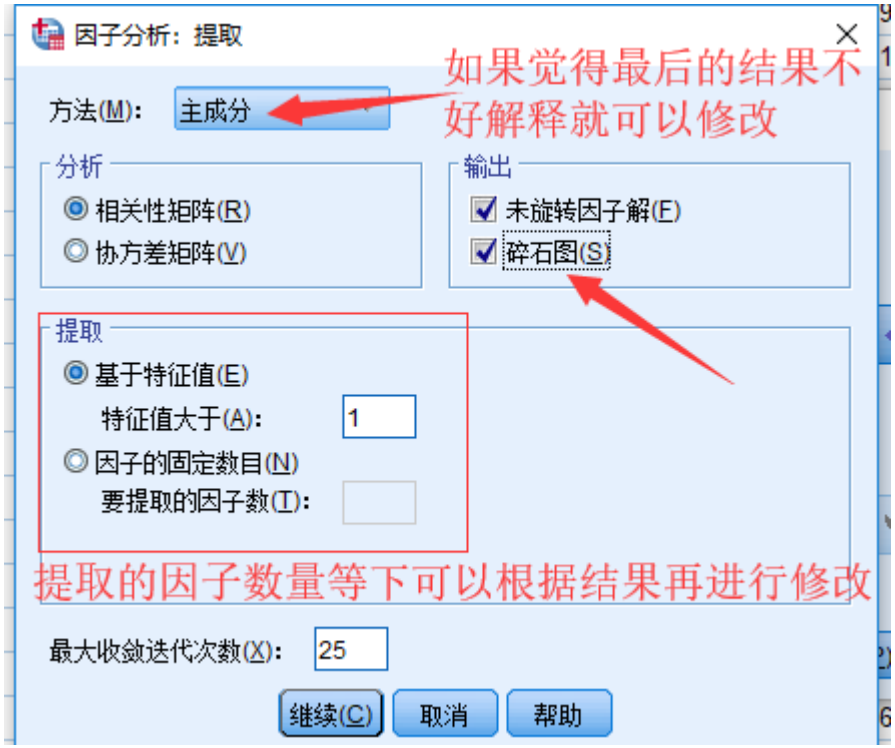
其中，Kaiser给出一个KMO检验标准：KMO>0.9,非常适合；0.8<KMO<0.9,适合；0.7<KMO<0.8, 一般；0.6<KMO<0.7,不太适合；KMO<0.5,不适合。

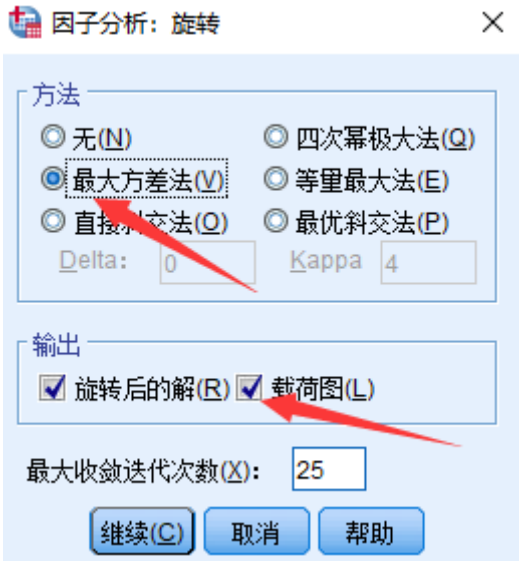
巴特利特球形检验

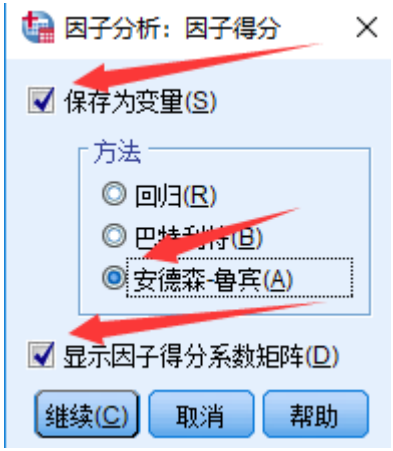
巴特利特球形检验是一种检验各个变量之间相关性程度的检验方法。一般在做因子分析之前都要进行巴特利特球形检验，用于判断变量是否适合用于做因子分析。巴特利特球形检验是以变量的相关系数矩阵为出发点的。它的原假设是相关系数矩阵是一个单位阵（不适合做因子分析，指标之间的相关性太差，不适合降维），即相关系数矩阵对角线上的所有元素都是1，所有非对角线上的元素都为0。巴特利特球形检验的统计量是根据相关系数矩阵的行列式得到的。如果该值较大，且**其对应的p值小于用户心中的显著性水平（一般为0.05），那么应该拒绝原假设，认为相关系数不可能是单位阵，即原始变量之间存在相关性，适合于作因子分析。**相反不适合作因子分析。

注意：用SPSS做因子分析时，在查看器中若得不到KMO检验和Bartlett检验结果，则说明你的样本量小于指标数了，需要增加样本量或者减少指标个数再来进行因子分析。

操作步骤







结果分析

注意：第一次运行因子分析的结果一般作为参考，首先我们要确定原始数据是否适合进行因子分析，即能否通过KMO检验和巴特利特球形检验。

KMO 和巴特利特检验

KMO 取样适切性量数。		.909
巴特利特球形度检验	近似卡方	719.113
	自由度	28
	显著性	.000

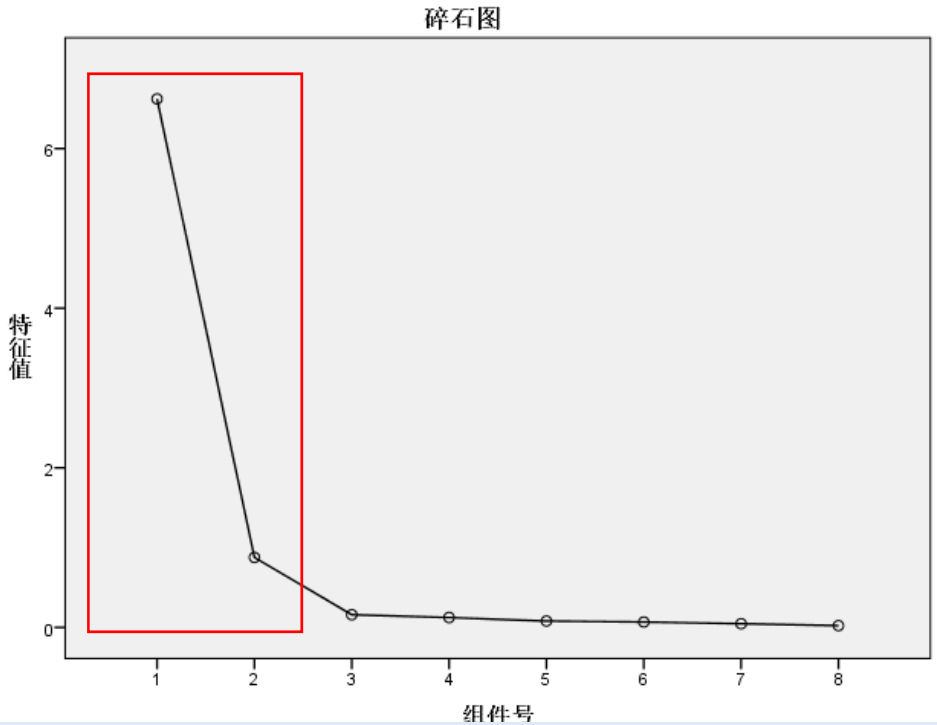
KMO检验标准：KMO>0.9,非常适合因子分析；0.8<KMO<0.9,适合； 0.7<KMO<0.8,一般；0.6<KMO<0.7,不太适合；KMO<0.5,不适合。

巴特利特球形检验：如果其统计量对应的p值小于用户心中的显著性水平（一般取0.05），那么应该拒绝原假设，认为相关系数不可能是单位阵，即原始变量之间存在相关性，适合于作因子分析；否则不适合作因子分析。

（1）KMO值等于0.909，说明数据适合进行因子分析；
（2）巴特利特球形检验的p值等于0.000，小于0.05，说明我们在95%的置信水平下拒绝原假设，即我们认为数据适合进行因子分析。

确定因子的数目

注意：第一次运行因子分析的结果一般作为参考，下面我们需要根据第一次运行的结果来确定公共因子的个数。

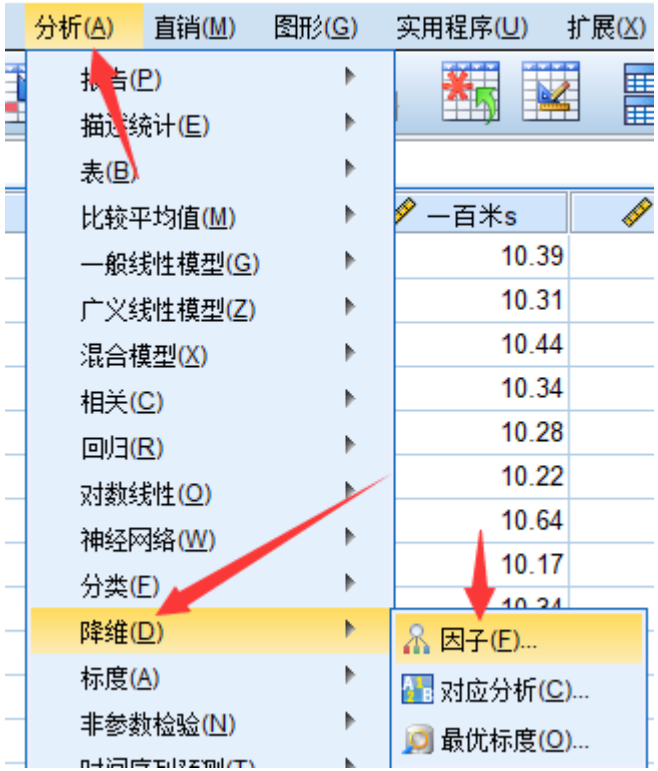


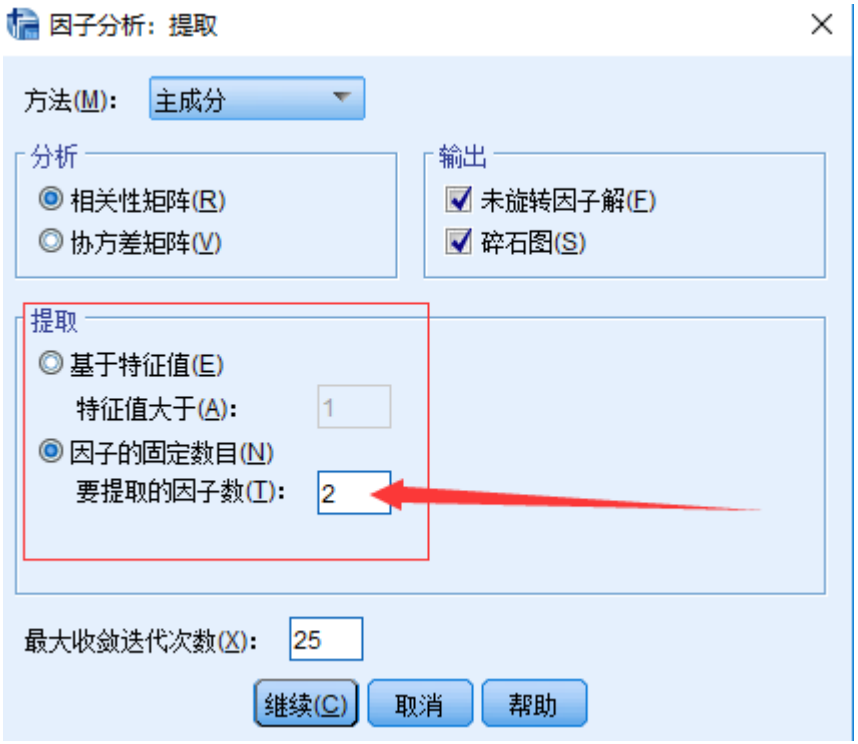
碎石检验（scree test）是根据碎石图来决定因素数的方法。Kaiser提出，可通过直接观察特征值的变化来决定因素数。当某个特征值较前一特征值的值出现较大的下降，而这个特征值较小，其后面的特征值变化不大，说明添加相应于该特征值的因素只能增加很少的信息，所以前几个特征值就是应抽取的公共因子数。

从碎石图可以看出，前两个因子对应的特征值的变化较为陡峭，从第三个因子开始，特征值的变化较为平坦，因此我们应选择两个因子进行分析。

（SPSS中文版翻译成了组件，实际应翻译为因子）

调整因子个数重新计算





这里选择的因子数就是刚刚我们通过碎石图得到的因子数。
（注意：碎石图得到的因子数只起到参考作用；在因子分析应用于某些专业问题上时，可能事先我们已经知道了最后要确定的因子数，这时候碎石图的意义就不大了）

对因子分析结果的介绍

（2）A的行元素平方和$h_i^2=\sum_{j=1}^{m}a_{ij}^2$：原始变量$x_i$对公因子依赖的程度

可以证明：$Var(x_i)=h_i^2+\sigma_i^2\ (i=1,2,\cdots,p)$

h_i^2反应了公因子对于x_i的影响，可以看成是公因子对于x_i的方差贡献，称为共性方差；而σ_i^2是特殊因子ε_i对x_i的方差贡献，称为个性方差。如果x经过了标准化，则$h_i^2+\sigma_i^2=1$.

公因子方差

	初始	提取
100米(s)	1.000	.950
200米(s)	1.000	.939
400米(s)	1.000	.892
800米(min)	1.000	.900
1500米(min)	1.000	.938
5000米(min)	1.000	.965
10000米(min)	1.000	.973
马拉松(min)	1.000	.943

提取方法：主成分分析法。

在论文中怎么解释这个值？
100米(s)这个变量的公因子方差为0.95，这可以解释为我们提取的两个公共因子对100米(s)这个变量的方差贡献率为95%，即这两个公共因子能够反映出（或者说保留）100米(s)这个变量95%的信息。

共性方差在SPSS中被称为了公因子方差。

注：正交旋转不会改变公因子方差。

总方差解释表

总方差解释

成分	初始特征值			提取载荷平方和			旋转载荷平方和		
	总计	方差百分比	累积 %	总计	方差百分比	累积 %	总计	方差百分比	累积 %
1	6.622	82.777	82.777	6.622	82.777	82.777	4.186	52.323	52.323
2	.878	10.970	93.747	.878	10.970	93.747	3.314	41.424	93.747
3	.159	1.992	95.739						
4	.124	1.551	97.289						
5	.080	.999	98.288						
6	.068	.850	99.137						
7	.046	.580	99.717						
8	.023	.283	100.000						

提取方法：主成分分析法。

上表为总方差解释表，给出了每个公共因子所解释的方差及累计和。
从“初始特征值”一栏中可以看出，前2个公共因子解释的累计方差达93.747%，而后面的公共因子的特征值较小，对解释原有变量的贡献越来越小，因此提取两个公共因子是合适的。
“提取载荷平方和” 一栏是在未旋转时被提取的2个公共因子的方差贡献信息，其与“初始特征值”栏的前两行取值一样。
“旋转载荷平方和”是旋转后得到的新公共因子的方差贡献信息，和未旋转的贡献信息相比，每个公共因子的方差贡献率有变化，但最终的累计方差贡献率不变，

成分矩阵

成分矩阵[a]

	成分 1	成分 2
100米(s)	.817	.531
200米(s)	.867	.432
400米(s)	.915	.233
800米(min)	.949	.012
1500米(min)	.959	-.131
5000米(min)	.938	-.292
10000米(min)	.944	-.287
马拉松(min)	.880	-.411

提取方法：主成分分析法。

a. 提取了 2 个成分。

旋转后的成分矩阵[a]

	成分 1	成分 2
100米(s)	.274	.935
200米(s)	.376	.893
400米(s)	.543	.773
800米(min)	.712	.627
1500米(min)	.813	.525
5000米(min)	.902	.389
10000米(min)	.903	.397
马拉松(min)	.936	.261

提取方法：主成分分析法。
旋转方法：凯撒正态化最大方差法。

a. 旋转在 3 次迭代后已收敛。

上面的“成分矩阵”是未经旋转的因子载荷矩阵，下面的“旋转后的成分矩阵”是经过旋转后的因子载荷矩阵。

观察两个表格可以发现，旋转后的每个公共因子上的载荷分配更清晰了，因而比未旋转时更容易解释各因子的意义。

我们在实际应用中只用关注旋转后的因子载荷矩阵即可。

因子载荷是变量与公共因子的相关系数，当某变量在某公共因子中的载荷绝对值越大，表明该变量与该公共因子更密切，即该公共因子更能代表该变量。

由此可知，本例中的第1个公共因子更能代表后面五个变量，我们可以称为长跑因子（或耐力因子）；

第2个公共因子更能代表前三个变量，我们可称为短跑因子（爆发力因子）。

旋转后的因子载荷散点图

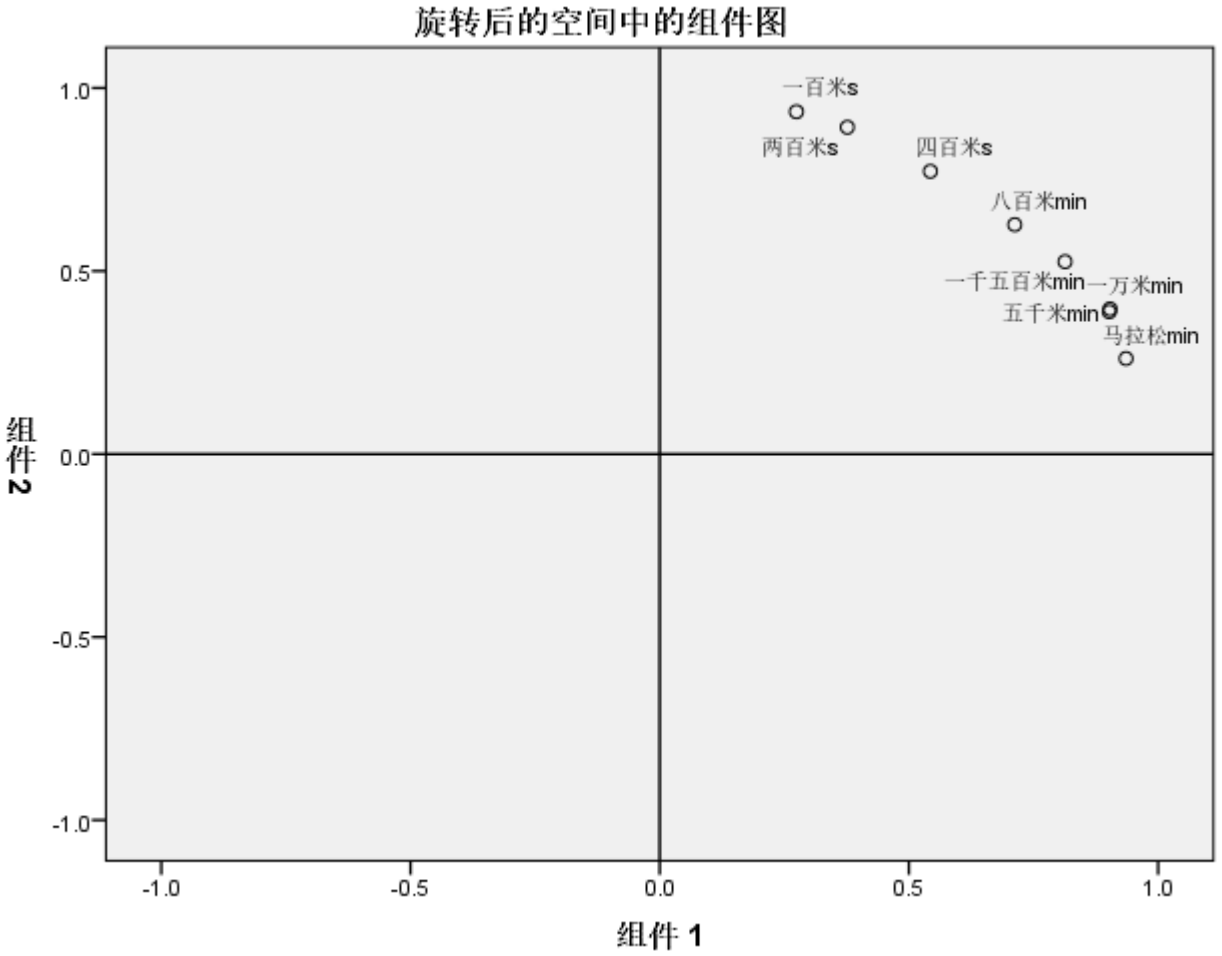
（SPSS中文版翻译成了组件，实际应翻译为因子）

旋转后的成分矩阵[a]

	成分 1	成分 2
100米(s)	.274	.935
200米(s)	.376	.893
400米(s)	.543	.773
800米(min)	.712	.627
1500米(min)	.813	.525
5000米(min)	.902	.389
10000米(min)	.903	.397
马拉松(min)	.936	.261

提取方法：主成分分析法。
旋转方法：凯撒正态化最大方差法。

a. 旋转在 3 次迭代后已收敛。



根据“旋转后的成分矩阵”的两列数据所作，由此图观察所得信息与从“旋转成分矩阵”所得信息一致。（如果有三个因子，那么画出来的图就是三维图）

因子得分

成分得分系数矩阵

	成分 1	成分 2
100米(s)	-.300	.540
200米(s)	-.222	.459
400米(s)	-.068	.291
800米(min)	.100	.103
1500米(min)	.207	-.019
5000米(min)	.324	-.161
10000米(min)	.321	-.156
马拉松(min)	.406	-.269

提取方法：主成分分析法。
旋转方法：凯撒正态化最大方差法。
组件得分。

min	五千米min	一万米min	马拉松min	FAC1_1	FAC2_1	变量
3.70	14.04	29.36	137.72	.31698	-.21309	
3.57	13.28	27.66	128.30	-.57116	-.79421	
3.60	13.26	27.72	135.90	-.57674	.19002	
3.60	13.22	27.45	129.95	-.78406	-.30439	
3.75	14.68	30.55	146.62	1.24631	-1.24424	
3.66	13.62	28.62	133.13	.01382	-.91372	
3.85	14.45	30.28	139.95	.40270	.70737	
3.63	13.55	28.09	130.15	-.16717	-.84726	
3.71	13.61	29.30	134.03	.02763	-.25960	
3.73	13.90	29.13	133.53	-.12811	.39332	
3.74	13.49	27.88	131.35	-.46260	.30630	
4.24	16.70	35.38	164.70	2.06399	3.89349	
3.84	14.03	28.81	136.58	-.48409	1.93453	
3.58	13.42	28.19	134.32	-.38475	-.37067	
3.61	13.50	28.11	130.78	-.59753	.03201	
3.82	14.91	31.45	154.12	2.20817	-1.54986	
3.61	13.27	27.52	130.87	-.78317	-.09703	
3.57	13.34	27.97	132.30	-.29090	-.95701	
3.56	13.17	27.42	129.92	-.54785	-.90713	
3.53	13.21	27.61	132.23	-.46683	-.97926	
3.51	13.01	27.51	129.13	-.69795	-.99127	
3.64	14.59	28.45	134.60	.30640	-.58354	

$$f_1 = -0.3*\widetilde{100米} - 0.222*\widetilde{200米} + \cdots + 0.406*\widetilde{马拉松}$$

$$f_2 = 0.54*\widetilde{100米} + 0.459*\widetilde{200米} + \cdots - 0.269*\widetilde{马拉松}$$

（这里加上弧形表示是标准化的原始变量）

和主成分分析一样，我们可以用因子得分f1和f2作为两个新的变量，来进行后续的建模（例如聚类、回归等）

注意：因子分析模型不能用于综合评价，尽管有很多论文是这样写的，但这是存在很大的问题的。例如变量的类型、选择因子的方法、旋转对最终的影响都是很难说清的。